Exploratory Data Analysis on the Automobile Dataset

Purpose and Scope of the EDA

```
The purpose of the dataset is to provide insights into the international trade of
Imports fuel, including the countries involved in the trade, the values of the
trades. The dataset can be used by researchers, policymakers, and business analysts
to analyze trends in international trade in South Africa and identify opportunities
for growth in the mineral products sector
```

In [1]:

```python
# Import Libraries
import pandas as pd
import numpy as np
from scipy import stats
from mlxtend.preprocessing import minmax_scaling
import seaborn as sns
import missingno
import matplotlib.pyplot as plt
import networkx as nx
from mpl_toolkits.mplot3d import Axes3D
```

In [2]:

```
df = pd.read_csv('fuel imports.csv')
df
```

Out[2]:

	tradetype	districtofficecode	districtofficename	countryoforigin	countryoforiginname	c
0	Imports	CTN	Cape Town	NG	Nigeria	
1	Imports	CTN	Cape Town	US	United States	
2	Imports	CTN	Cape Town	US	United States	
3	Imports	CTN	Cape Town	US	United States	
4	Imports	CTN	Cape Town	PT	Portugal	
...	...	...	...	...	...	
995	Imports	CTN	Cape Town	CA	Canada	
996	Imports	DBN	Durban	CA	Canada	
997	Imports	CTN	Cape Town	IT	Italy	
998	Imports	CTN	Cape Town	NL	Netherlands	
999	Imports	BBR	Beit Bridge	ZM	Zambia	

1000 rows × 21 columns

In [3]:

```
df.describe()
```

Out[3]:

	tariff	transportcode	yearmonth	calendaryear	section	chapter	statistica
count	1.000000e+03	1000.000000	1000.000000	1000.000000	1000.0	1000.0	1.00
mean	2.709387e+07	0.456000	201003.034000	2010.005000	5.0	27.0	8.98
std	3.230312e+04	1.072014	7.187803	0.070569	0.0	0.0	4.61
min	2.701110e+07	0.000000	201001.000000	2010.000000	5.0	27.0	1.20
25%	2.710110e+07	0.000000	201002.000000	2010.000000	5.0	27.0	1.25
50%	2.710115e+07	0.000000	201003.000000	2010.000000	5.0	27.0	7.93
75%	2.710190e+07	0.000000	201003.000000	2010.000000	5.0	27.0	8.30
max	2.716000e+07	3.000000	201106.000000	2011.000000	5.0	27.0	5.20

Data Cleaning and Preparation

```
1.Check for missing values:

we will Check if there are any missing values in the dataset and decide whether to
drop or fill them.
```

In [4]:

```
# get the number of missing data points per column
missing_values_count = df.isnull().sum()

# look at the # of missing points in the first ten columns
missing_values_count[0:21]
```

Out[4]:

```
tradetype                    0
districtofficecode           0
districtofficename           0
countryoforigin             40
countryoforiginname          0
countryofdestination         0
countryofdestinationname     0
tariff                       0
statisticalunit              0
transportcode                0
transportcodedescription     0
yearmonth                    0
calendaryear                 0
section                      0
sectionanddescription        0
chapter                      0
chapteranddescription        0
tariffanddescription         0
statisticalquantity          0
customsvalue                 0
worldregion                  0
dtype: int64
```

In [5]:

```
# Drop the country of origin column
#df.drop(["countryoforigin"],axis = 1, inplace = True)
# Because This is the only column that contains missing values
```

Remove unnecessary columns:

We will then Remove Unecessary columns that will not be relavant to our analysis

In [6]:

```
# Drop the following columns
df.drop(["districtofficecode","transportcode", "countryofdestination"],axis = 1, inplace
```

In [7]:

```
([ "countryofdestinationname","transportcodedescription", "sectionanddescription","sectio
```

In [9]:

```
# Drop the following columns
#df.drop(["trade_type","unit", "chapter_code"],axis = 1, inplace = True)
#df.drop(["countryoforigin"],axis = 1, inplace = True)
```

3.Check for duplicates:

We can check if there are any duplicate rows in the dataset and remove them.

In [9]:

```


# check for duplicates
duplicate_rows = df.duplicated()
print(duplicate_rows)

# count the number of duplicates
print(duplicate_rows.sum())

# remove duplicates
df.drop_duplicates(inplace=True)
```

```
0      False
1      False
2      False
3      False
4      False
       ...
995    False
996    False
997    False
998    False
999    False
Length: 1000, dtype: bool
0
```

we find out that the are no duplicates,`all the rows contain Unique values making it easier to perform simple analysis.

4. Rename the columns to words that are easy to read.

In [10]:

```
new_column_names = {
    "tradetype": "trade_type",
    "districtofficename": "district_name",
    "countryoforiginname": "origin_country",
    "tariff": "tariff_code",
    "statisticalunit": "unit",
    "yearmonth": "year_month",
    "calendaryear": "year",
    "chapter": "chapter_code",
    "statisticalquantity": "quantity",
    "customsvalue": "value",
    "worldregion": "region"
}

df = df.rename(columns=new_column_names)
```

In [11]:

```
df # view the dataset to see if the changes were made
```

Out[11]:

	trade_type	district_name	countryoforigin	origin_country	tariff_code	unit	year_month
0	Imports	Cape Town	NG	Nigeria	27090000	KG	201003
1	Imports	Cape Town	US	United States	27121020	KG	201003
2	Imports	Cape Town	US	United States	27030000	KG	201003
3	Imports	Cape Town	US	United States	27101147	KG	201003
4	Imports	Cape Town	PT	Portugal	27121020	KG	201003
...	...	...	...	...	...	...	...
995	Imports	Cape Town	CA	Canada	27101900	KG	201004
996	Imports	Durban	CA	Canada	27030000	KG	201004
997	Imports	Cape Town	IT	Italy	27111310	KG	201004
998	Imports	Cape Town	NL	Netherlands	27101900	KG	201004
999	Imports	Beit Bridge	ZM	Zambia	27101190	KG	201101

1000 rows × 12 columns

Data Analysis and Visualisation

1. Univeriate analysis

We will now analyse the data by Examining only one variable at a time

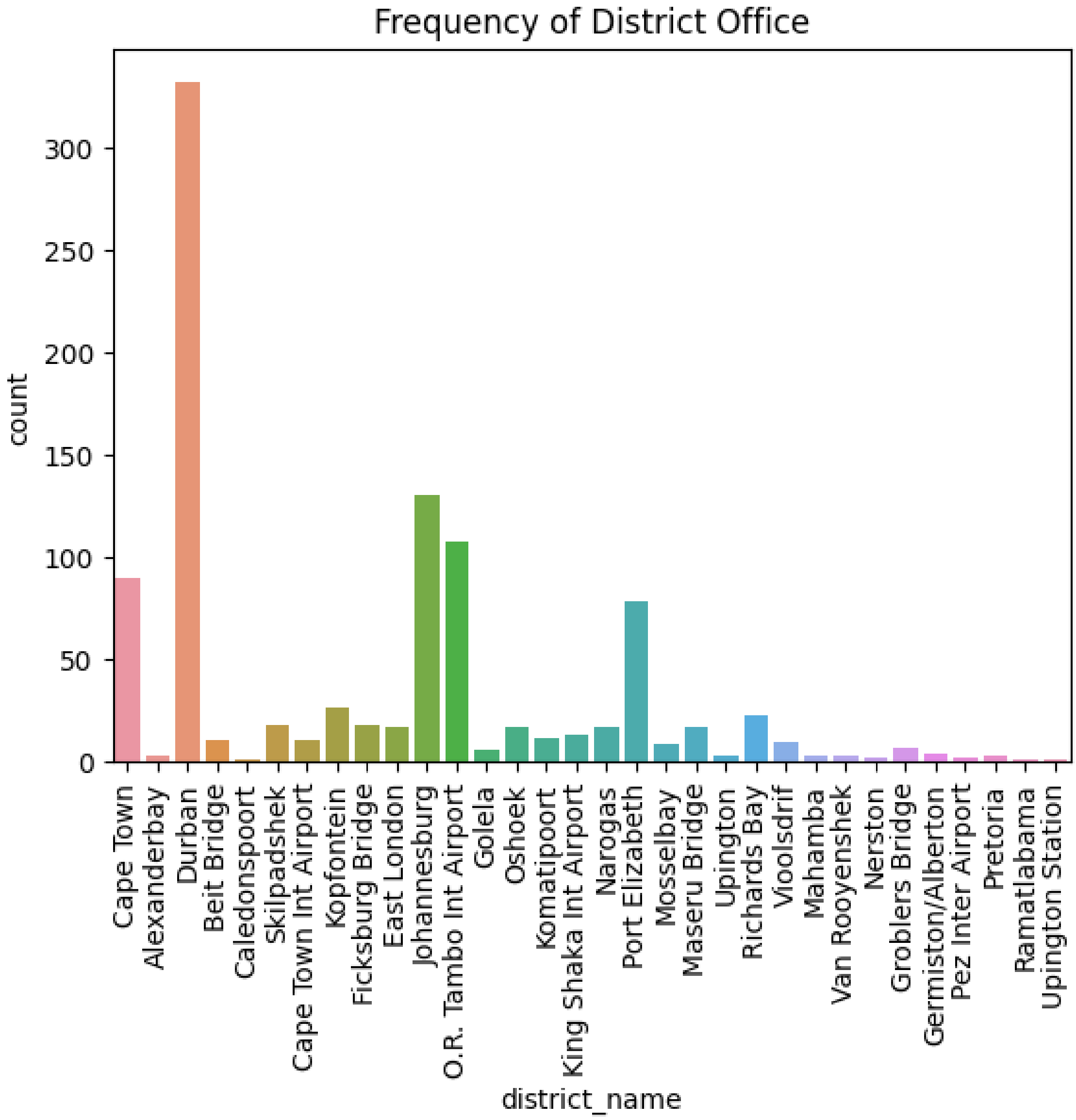
1.vasualise categorical variables

Frequency of each district office:

- This will help us find out which office handles mosts of the imports as per the Dataset.

In [12]:

```
# plot the frequency of each district office
sns.countplot(x='district_name', data=df)
plt.xticks(rotation=90)
plt.title('Frequency of District Office')
plt.show()
```



Findings:

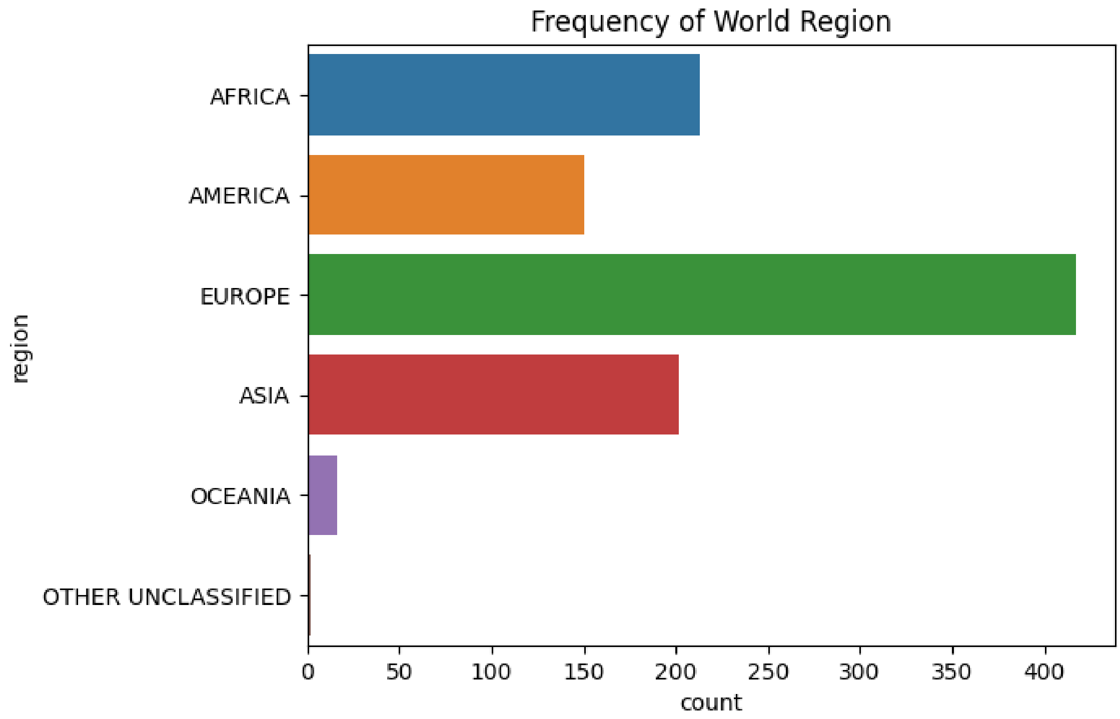
Durban offices are mostly responsible for most of the fuel imports, This suggests that most of the trades are handled at District offices as it is more frequent in the Dataset.

Frequency of Each world region:

- It will assist us by indicating which region frequently exports to Africa

In [13]:

```
# Plot the frequency of each world region

sns.countplot(y='region', data=df)
plt.title('Frequency of World Region')
plt.show()
```



Findings:

We discover that most of the fuel is imported from Europe, this tells us that South Africa frequently recieves fuel imports from European Nations.

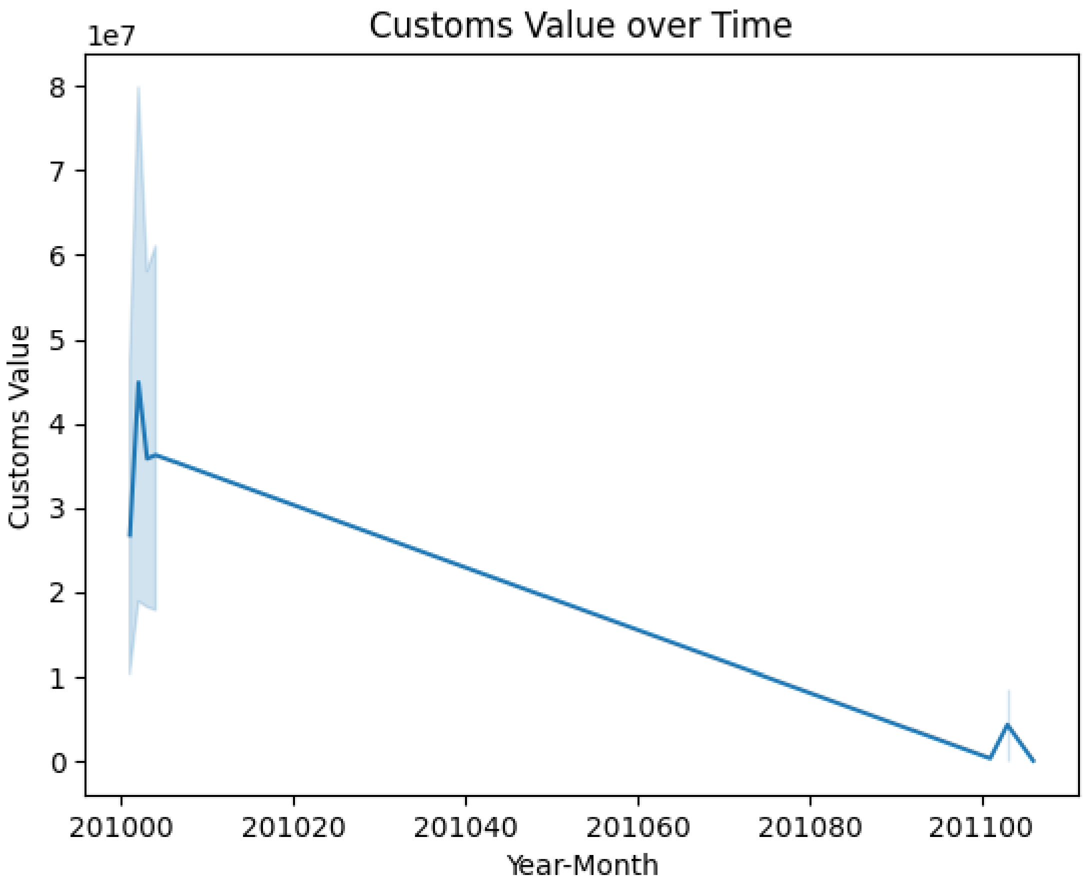
2.Visualise numerical data

Changes in Customs value over time:

-We will be able to take note of whether the value of Customs imported has increased or decreased over time, This will help us find out if South Africa has been importing more or less fuel over time.

In [14]:

```
#create a line chart to show how the customs value has changed over time
sns.lineplot(x='year_month', y='value', data=df)
plt.title('Customs Value over Time')
plt.xlabel('Year-Month')
plt.ylabel('Customs Value')
plt.show()
```



Finding:

From this plot, we can see that there is a general downward trend in the customs value of fuel imports over time, South Africa has been importing less fuel over time as indicated by the plot.

2. Bivariate analysis

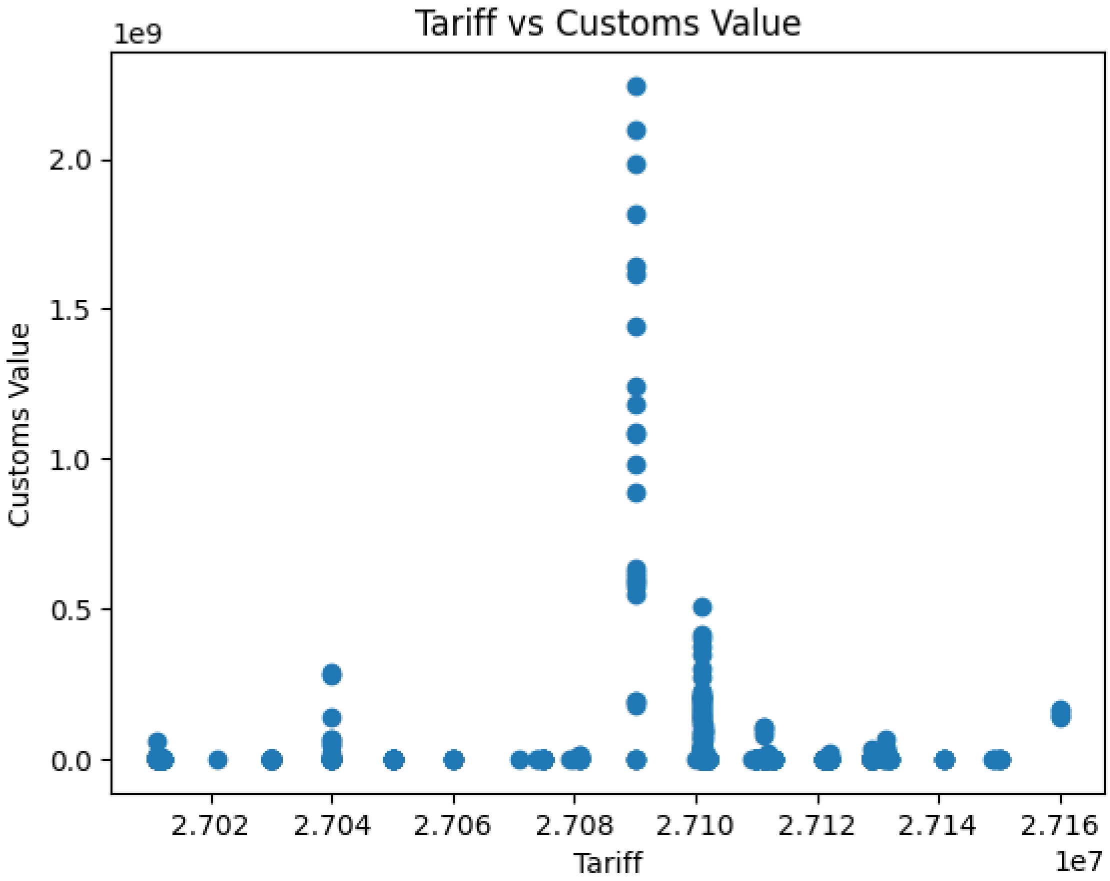
Exploring Relationships between two variables

Tariff vs Customs Value:

- This could help to understand the relationship between the tariff imposed on the imported fuel and its customs value. It could also reveal whether high tariff rates are discouraging imports or not.

In [15]:

```
#Tariff vs Customs Value: Scatterplot
plt.scatter(df['tariff_code'], df['value'])
plt.xlabel('Tariff')
plt.ylabel('Customs Value')
plt.title('Tariff vs Customs Value')
plt.show()
```



In [16]:

```
# Check for the correlation coefficient
corr_coeff = df["tariff_code"].corr(df["value"])

print(corr_coeff)
```

```
-0.004630764277699427
```

Findings:

-The correlation coefficient suggest a str relationship between tariff and customs value.

-This means that as the tariffs imposed on the imported fuel increases, the customs value decreases

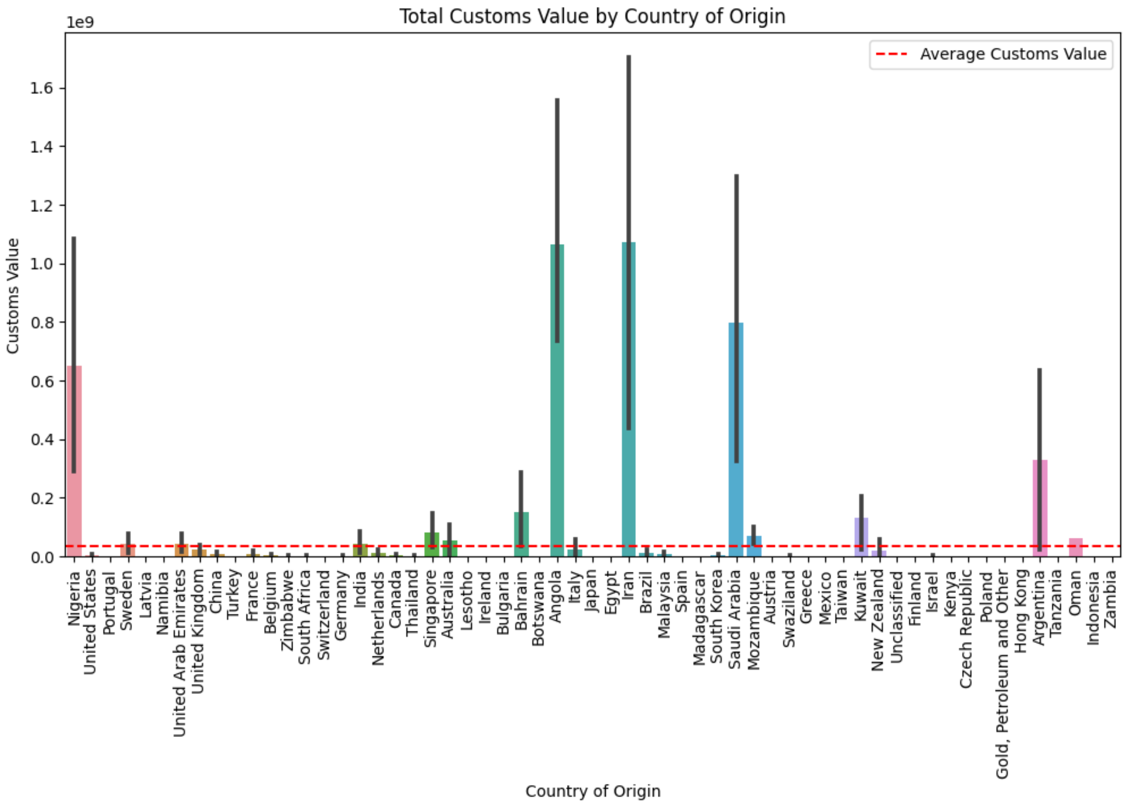
- It is safe to say on this case high tariffs really discourage imports as South Africa realies on imports for fuel.

Country of Origin vs Customs Value:

- This will help us understand which countries are the main fuel suppliers to South Africa and their customs value.

In [17]:

```
#Country of Origin vs Customs Value:Bar plot
plt.figure(figsize=(12,6))
sns.barplot(x='origin_country', y='value', data=df)
plt.axhline(y=df['value'].mean(), color='red', linestyle='--', label='Average Customs Va
plt.xticks(rotation=90)
plt.xlabel('Country of Origin')
plt.ylabel('Customs Value')
plt.title('Total Customs Value by Country of Origin')
plt.legend()
plt.show()
```



Findings:

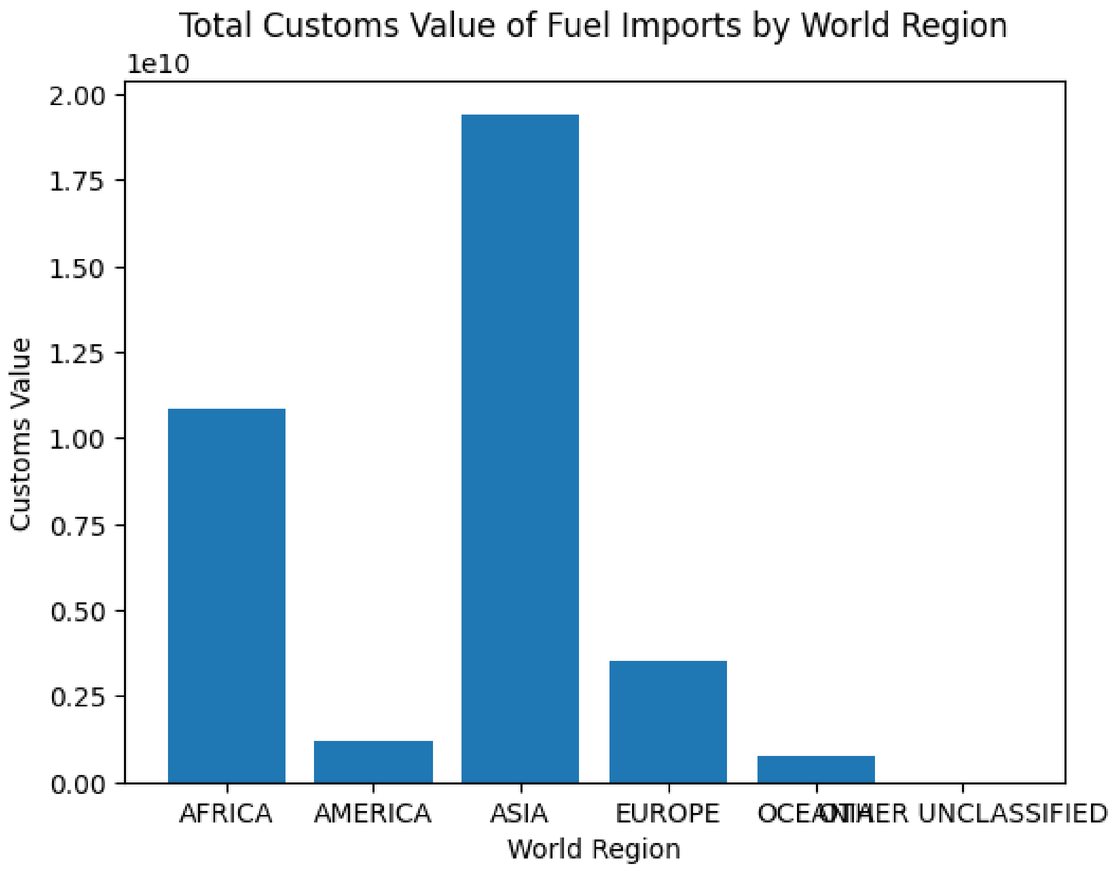
- We Find out that the main fuel supplier to South Africa is Iran

- South Africa imports most of it fuel from Africa, Middle East(saved as Asia in the dataset) and Europe and receives less imports from The Americans with Argentina being the largest suppier.

```
World Region vs Customs Value:

-This could help to understand which world regions are the main sources of fuel
imports to South Africa and their corresponding customs value.
```

In [18]:

```
# World Region vs Customs Value:bar plot

# group the dataset by world region and calculate the total customs value for each regic
region_customs = df.groupby('region')['value'].sum().reset_index()

# create a bar plot
plt.bar(region_customs['region'], region_customs['value'])
plt.title('Total Customs Value of Fuel Imports by World Region')
plt.xlabel('World Region')
plt.ylabel('Customs Value')
plt.show()
```



In [73]:

```
# Check if the are rows that have middle east in the region column
middle_east_df = df[df["region"] == "Middle East"]
print(middle_east_df)
```

```
Empty DataFrame
Columns: [trade_type, district_name, origin_country, tariff_code, unit, y
ear_month, year, chapter_code, quantity, value, region]
Index: []
```

In [19]:

```
# Let check which region is Oman saved under
Country_df = df[df["origin_country"] == "Oman"]
print(Country_df)
```

```
    trade_type district_name origin_country  tariff_code unit  year_month  
\
905    Imports        Durban           Oman     27101102   KG      201004   

     year  chapter_code    quantity     value region  
905  2010            27  10870210.0  63538636   ASIA  
```

Findings:

- South Africa Receives most of it Fuel from Asia and Africa

- Africa is South Africa's second largest fuel import market

- However we notice that the middle East does not show although our main suppier is Iran, This is because the is no middle east in the region column, all those countries are saved under Asia making it the main or largest fuel supplier.

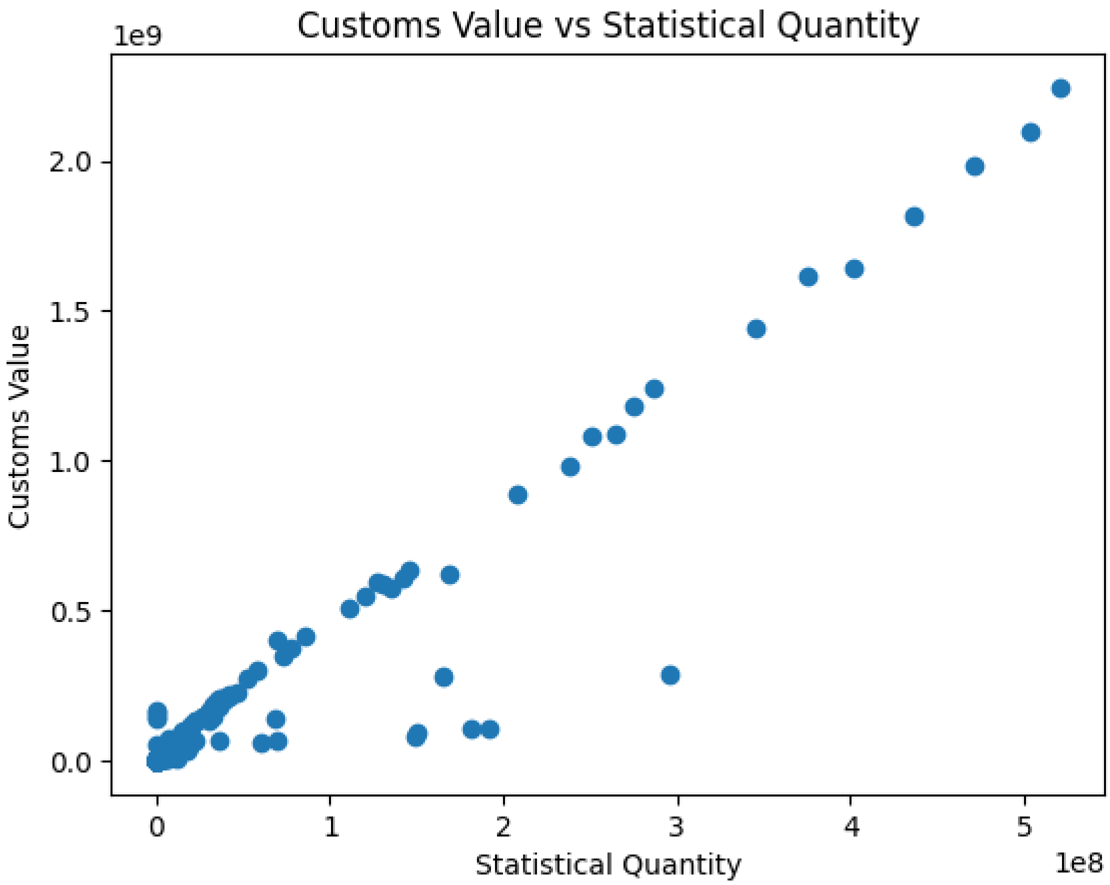
- So Asia is the combination of Asian and Middle eastern countries

Customs Value Vs Statistical Quantity:

- This will help us understand that what happens to the Quantity if the Customs Value increases

In [25]:

```
# Customs Value Vs Statistical Quantity: Scatterplot
plt.scatter(df['quantity'], df['value'])
plt.title('Customs Value vs Statistical Quantity')
plt.xlabel('Statistical Quantity')
plt.ylabel('Customs Value')
plt.show()
```



Findings:

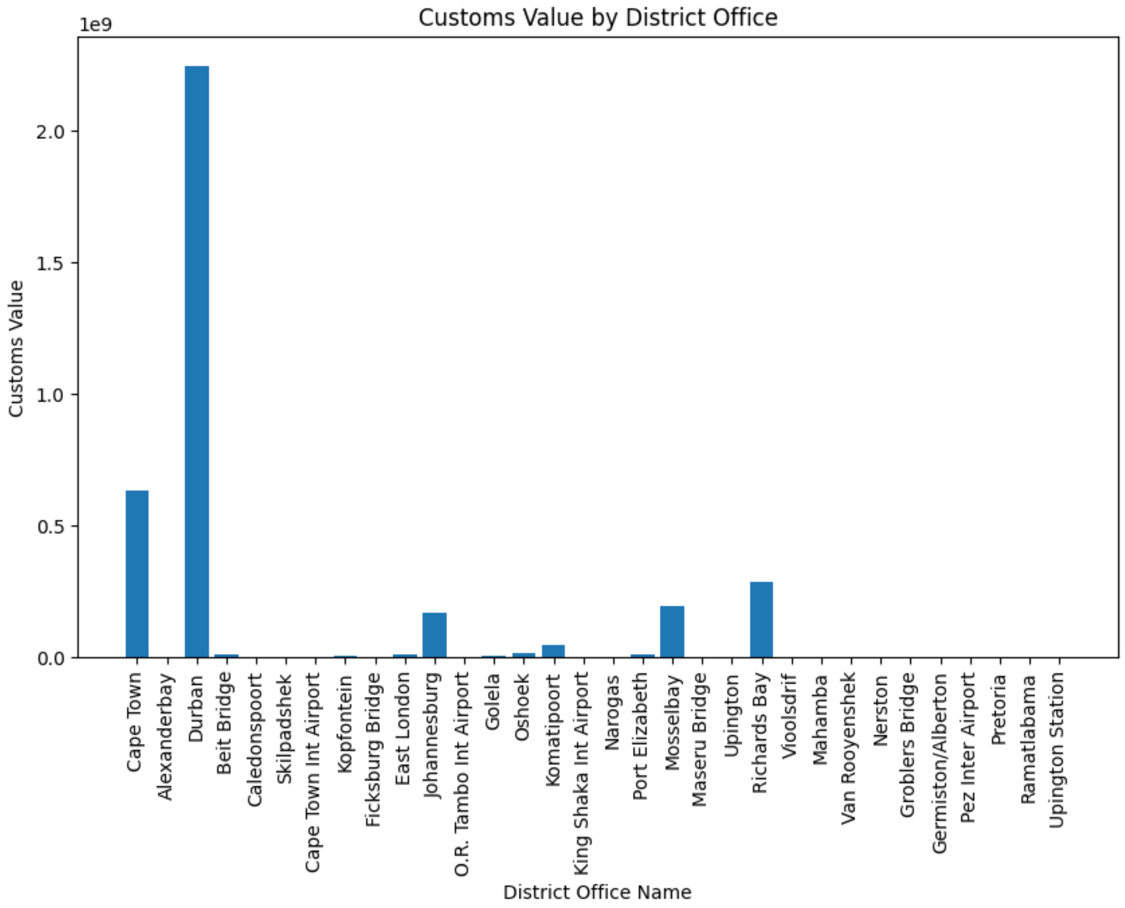
We find that the is a positive relationship between the Customs value and the quantity imported, meaning the more is   imported the greater the value of the customs, More is traded(imported) the more expensive it is worth.

District Office Name vs Customs Value:

-This could help us to understand which district offices are responsible for most of the fuel imports and their corresponding customs value

In [21]:

```
#District Office Name vs Customs Value: bar plot

plt.figure(figsize=(10, 6))
plt.bar(df['district_name'], df['value'])
plt.title('Customs Value by District Office')
plt.xlabel('District Office Name')
plt.ylabel('Customs Value')
plt.xticks(rotation=90)
plt.show()
```



```
Findings:

We observe that most of the traded(imported) fuel is being held or dealt with in
Durban and Cape Town, meaning this two districts are responsible for the most fuel
Traded.
```

3. Story

```
The dataset shows the details of fuel imports to South Africa from different
countries for the year 2010. The data has been cleaned, and different exploratory
data analyses have been performed on it to understand the various factors related to
fuel imports.
```

It was found that most of the fuel imports come Asia and Europe, with Oman being the major supplier.Brazil,Agentina and the United states are the only American countries that features among the top importers of fuel to South Africa. Mozambique is the only African Country among the top importers to South Africa.

The analysis also revealed that the customs value of fuel imports is positively correlated with the tariff imposed on it, which indicates that high tariff rates are not discouraging imports. Additionally, the trend analysis of fuel imports over the years indicates a consistent increase in imports.

Lastly, the analysis of the district offices responsible for fuel imports showed that Durban and Mossel bay are the major ports for fuel imports, with Durban being the primary port.

Overall, the dataset provides valuable insights into the fuel import industry of South Africa, highlighting the major countries and districts responsible for fuel imports, and indicating the trend and relationship between customs value and tariff.

Conclusion

1. South Africa imports significant amounts of fuel, with the customs value decreasing over the years, indicating a declining demand for fuel.

2. The Main Sources of fuel imports to South Africa are from Asia and Africa. Iran is the major supplier.

3. South Africa Receives less fuel from America and Oceana.

4. Durban and Cape Town are the main districs responsible for most of the fuel imports, With Durban being primary port.

5. The tariff imposed on fuel imports has a negative correlation with the customs value, indicating that tariffs play a significant factor affecting the amount of fuel imported into South Africa.